Active Learning for Black-Box Adversarial Attacks in EEG-Based Brain-Computer Interfaces

Xue Jiang, Xiao Zhang and Dongrui Wu
School of Artificial Intelligence and Automation
Huazhong University of Science and Technology, Wuhan, China
Email: xuejiang@hust.edu.cn, xiao_zhang@hust.edu.cn, dwu@hust.edu.cn

Abstract—Deep learning has made significant breakthroughs in many fields, including electroencephalogram (EEG) based brain-computer interfaces (BCIs). However, deep learning models are vulnerable to adversarial attacks, in which deliberately designed small perturbations are added to the benign input samples to fool the deep learning model and degrade its performance. This paper considers transferability-based black-box attacks, where the attacker trains a substitute model to approximate the target model, and then generates adversarial examples from the substitute model to attack the target model. Learning a good substitute model is critical to the success of these attacks, but it requires a large number of queries to the target model. We propose a novel framework which uses query synthesis based active learning to improve the query efficiency in training the substitute model. Experiments on three convolutional neural network (CNN) classifiers and three EEG datasets demonstrated that our method can improve the attack success rate with the same number of queries, or, in other words, our method requires fewer queries to achieve a desired attack performance. To our knowledge, this is the first work that integrates active learning and adversarial attacks for EEG-based BCIs.

Keywords—brain-computer interfaces; adversarial examples; active learning; black-box attack

I. INTRODUCTION

A brain-computer interface (BCI) is a communication system that connects a human brain and a computer [1]. Direct dialogue between the brain and the computer can be achieved by the established information path. Electroencephalogram (EEG) is the most frequently used input signal in BCIs, due to its low cost and convenience [2]. Various paradigms are used in EEG-based BCIs, such as P300 evoked potentials [3]–[6], motor imagery (MI) [7], steady-state visual evoked potential (SSVEP) [8], etc.

Deep learning has achieved great success in numerous fields. Multiple convolutional neural network (CNN) classifiers have also been proposed for EEG-based BCIs. Lawhern et al. [9] proposed EEGNet, which can be applied to different BCI paradigms. Schirrmeister et al. [10] designed a deep CNN model (DeepCNN) and a shallow CNN model (ShallowCNN). In addition, there were some studies to convert EEG signals into images and then classify them with deep learning models [11]–[13]. This paper considers only CNN models (i.e., EEGNet, DeepCNN, ShallowCNN) which take the raw EEG signals as the input.

Despite their state-of-the-art performance, recent studies have shown that deep learning models are vulnerable to adversarial examples, which are crafted by adding small imperceptible perturbations to benign examples to degrade the performance of a well-trained deep learning model. For example, in face recognition, an adversarial perturbation can be attached to the glasses, and the attacker who wears it can avoid being recognized, or be recognized as another person [14]. In image classification, the adversarial examples can fool a deep learning model to give incorrect image labels [15]–[18]. Many adversarial attacks have also been performed in speech recognition [19], malware classification [20], semantic segmentation [21], etc. Recently, Zhang and Wu [22] verified that adversarial examples exist in EEG-based BCIs and proposed several adversarial attack approaches.

Many effective algorithms for generating adversarial examples have been proposed, such as the fast gradient sign method (FGSM) [15], the C&W method [23], L-BFGS [16], the basic iterative method [17], DeepFool [24], etc. These methods mainly considered the white-box attack scenario, where the attacker has full access to the target model, including its architecture and parameters. Accordingly, the attacker can perform attacks by adding perturbations along the direction calculated by gradient-based strategies or optimization-based strategies. However, the white-box setting requires full information of the target model, making it impractical in many real-world applications.

In this paper, we focus on a more realistic and challenging black-box attack scenario, where the attacker can only observe the target model’s responses to inputs but has no information about its architecture, parameters, and training data. The attacker needs to generate adversarial examples whose perturbations are constrained to a magnitude threshold with limited queries. Papernot et al. [25] proposed a black-box attack method which generated adversarial examples for a white-box substitute model and attacked the black-box target model based on the transferability. Zhang and Wu [22] proposed an unsupervised fast gradient sign method (UFGSM) to craft adversarial examples for black-box attacks in EEG-based BCIs. However, these transferability-based approaches suffered from low query efficiency: they usually require a large number of queries to build a substitute model that is sufficiently similar to the target model.

To address this issue, we introduce a query synthesis based active learning strategy to transferability-based black-
black-box attacks of EEG-based BCIs. Given a small amount of
initial training EEG epochs for the substitute model, we first
randomly obtain a pair of opposite instances in each iteration,
i.e., two instances of different classes. According to the initial
opposite-pair, we use a binary search strategy to synthesize
another opposite-pair close to the current classification bound-
ary in the input space. After that, we synthesize queries along
the perpendicular line of the previously found opposite-pair.
This query synthesis based active learning strategy can directly
synthesize queries which are close to the target model’s
decision boundary and well scattered. It improves the query
efficiency by directly searching for informative examples in
the input space for substitute model training, instead of taking
a fixed step along the gradient direction. Experiments on
three CNN classifiers and three BCI datasets demonstrated the
effectiveness of our proposed method.

The remainder of this paper is organized as follows: Section
II introduces related work on black-box adversarial attacks
and active learning. Section III proposes our query synthesis
based active learning approach for crafting adversarial exam-
ples for black-box attacks in EEG-based BCIs. Section IV
evaluates the attack performance of our proposed approach.
Finally, Section V draws conclusion.

II. RELATED WORK

In this section, we briefly review previous studies on black-
box attacks and active learning.

A. Black-Box Attacks

Black-box attacks can be roughly divided into three cate-
gories: decision-based, score-based and transferability-based.
Decision-based attacks were first proposed by Brendel et
al. [26]. Its main idea is to gradually reduce the magnitude of
the adversarial perturbation while ensuring its effectiveness.
Score-based attacks rely on the model’s output scores, e.g.,
class probabilities or logits, to estimate the gradients and then
generate adversarial examples [27], [28]. Transferability-based
attacks were first proposed by Papernot et al. [25] in image
classification. The attacker trains a substitute model, which
solves the same classification problem as the target model, to
generate adversarial examples for the target model.

In transferability-based black-box attacks, the key step is to
learn a substitute model whose decision boundary resembles
the target model’s. Papernot et al. [25] used Jacobian-based
dataset augmentation to synthesize a substitute training set
and labeled it by querying the target model. By alternatively
augmenting the training set and updating the substitute model,
the CNN model. Formally, an adversarial example satisfies the following constraints:
\[ f_{\theta}(x^*_i) \neq y_i, \]  
\[ D(x_i, x^*_i) \leq \epsilon, \] where \( D(\cdot, \cdot) \) is a distance metric, and \( \epsilon \) constrains the magnitude of the adversarial perturbation. (1) ensures the success of
the attack, and (2) ensures that the perturbation is not larger than a predefined upper bound \( \epsilon \).

B. Active Learning

Active learning is an effective way to reduce the data
labeling effort, by actively selecting the most useful instances
to label. There are two main scenarios of active learning in the
literature: query synthesis [29]–[32] and sampling. The latter
can be further divided into stream-based sampling [33], [34]
and pool-based sampling [35]–[38]. Sampling-based active
learning selects real unlabeled instances from a pool or steam
for labeling. In query synthesis based active learning, one
can query any data instance in the input space, including
synthesized instances.

Intuitively, query synthesis can be applied to the training
process of the substitute model to improve the query efficiency
of black-box attacks, because it can actively synthesize more
informative EEG epochs than generating some epochs in
Jacobian-based way. Furthermore, compared with sampling,
it is more efficient to synthesize a query directly instead of
evaluating every instance in an unlabeled data pool. Here we
don’t need to worry about the fact that some synthesized EEG
epochs may be unrecognizable to human [39], because the
target model can label any instance in the input space. Our
ultimate goal is to identify the decision boundary of the target
model by a minimum number of queries.

III. METHODOLOGY

In this section, we first introduce the transferability-based
black-box attack setting for CNN classifiers in EEG-based B-
CIs, and then describe the query-synthesis-based active learn-
ing strategy for training the substitute model, and adversarial
example crafting for the target model.

A. Attack Setting

The attack framework in this paper is the same as our
previous work [22], where the attackers can add adversarial
perturbations before the machine learning modules.

Let \( x_i \in \mathbb{R}^{C \times T} \) be the \( i \)-th raw EEG epoch (\( i = 1, \ldots, n \)),
where \( C \) is the number of EEG channels and \( T \) the number of
the time domain samples. Let \( f_{\theta}(x_i) \rightarrow y_i \) denote the CNN
model that predicts the label for an input EEG epoch.

Given a target CNN model and a normal EEG epoch \( x_i \), the
task is to generate an adversarial example \( x^*_i \) misclassified by
the CNN model. Formally, an adversarial example \( x^*_i \) should
satisfy the following constraints:
\[ f_{\theta}(x^*_i) \neq y_i, \]  
\[ D(x_i, x^*_i) \leq \epsilon, \]
Next we describe an algorithm to learn a substitute model for a given target CNN classifier by querying it for labels on query-synthesized inputs, and then introduce the UFGSM approach [22] to craft adversarial examples on the trained substitute model, which can also be transferred to the target model. We start with binary classification, and then extend it to multi-class tasks.

B. Query-synthesis-based Dataset Augmentation for Training the Substitute Model

In the black-box attack scenario, we have no access to the architecture, parameters and training data of the target model, but can input EEG trials to the target model and observe the corresponding outputs to probe the model. As in [25], we also adopt the transferability-based approach to implement the black-box attack. The difference is that we use a query-synthesis-based active learning strategy as the data augmentation technique in substitute model training rather than the Jacobian-based approach. The oracle in query synthesis active learning is the target model.

1) Binary search synthesis: Assume a small labeled training set $S_0$ has been obtained from querying the target model. An initial substitute model $f'_0$ can be trained on this set. Suppose $\{x^+_0, x^-_0\}$ is an opposite-pair in $S_0$. Then, we query their middle point on the substitute model to find another opposite-pair closer to the decision boundary, using Algorithm 1.

Algorithm 1: Binary search synthesis. $\{x^+, x^-\} = BinarySearch(\{x^+_0, x^-_0\}, f', m)$

Input: $\{x^+_0, x^-_0\}$, initial opposite-pair of EEG epochs; $f'$, current substitute model; $m$, maximum number of binary search iterations.

Output: $\{x^+, x^-\}$, an opposite-pair of EEG epochs.

$x^+ = x^+_0$; $x^- = x^-_0$;

for $i = 1$ to $m$ do

$x_b = (x^+ + x^-)/2$;

Query $f'$ for $y_b$, the label of $x_b$;

if $y_b$ is positive then

$x^+ \leftarrow x_b$

else

$x^- \leftarrow x_b$

end

end

return $\{x^+, x^-\}$

2) Mid-perpendicular synthesis: There is an obvious limitation in binary search synthesis: if we always use binary search to generate training epochs, they may concentrate in one area and lack diversity. Therefore, we synthesize the next query along the mid-perpendicular direction after we find an opposite-pair close enough to the decision boundary. Specifically, we find the opposite-pair’s orthogonal vector by Gram-Schmidt process [32], set the magnitude of the orthogonal vector to $q$, then move it to a more precise midpoint. The details are shown in Algorithm 2.

Algorithm 2: Mid-perpendicular synthesis. $x_s = MidPerp(\{x^+_0, x^-_0\}, f', k, q)$

Input: $\{x^+_0, x^-_0\}$, an opposite-pair of EEG epochs; $f'$, current substitute model; $m$, maximum number of binary search iterations; $q$, magnitude of the orthogonal vector.

Output: $x_s$, an synthesized EEG epoch.

$x_1 = x^+_0 - x^-_0$;

Generate an EEG epoch $x_2$ randomly;

Find the orthogonal direction by Gram-Schmidt process:

$x_2 = q \cdot (x_2 - (x_1, x_2)/(x_1, x_2) \times x_1)$;

$\{x^+, x^-\} = BinarySearch(\{x^+_0, x^-_0\}, f', m)$;

$x_s = x_2 + (x^+ + x^-)/2$.

return $x_s$

With this query-syntheses-based active learning strategy, the entire substitute model training process for binary classification is shown in Algorithm 3.

Algorithm 3: Query-synthesis-based substitute model training strategy.

Input: $f$, the target model; $S_0$, a set of unlabeled EEG epochs; $N_{max}$, maximum number of training epochs; $n_{max}$, maximum number of synthesized EEG epochs in one iteration; $m$, maximum number of binary search iterations; $q$, magnitude of the orthogonal vector.

Output: $f'$, a trained substitute model

Label $S_0$ by querying $f$ to obtain an initial training set $D$;

Initialize $f'$ and pre-train $f'$ on $D$;

$\Delta D = \emptyset$;

for $N = 1$ to $N_{max}$ do

$\Delta S = \emptyset$;

for $n = 1$ to $n_{max}$ do

Select an opposite pair $x_0^+$ and $x_0^-$ randomly from $D$;

$\{x^+_0, x^-_0\} = BinarySearch(\{x^+_0, x^-_0\}, f', m)$;

$x_s = MidPerp(\{x^+_0, x^-_0\}, f', m, q)$;

$\Delta S \leftarrow \Delta S \cup \{x_s\}$;

end

$\Delta D = \{(x_i, f(x_i)) \}_{x_i \in \Delta S}$;

$D \leftarrow D \cup \Delta D$;

Train $f'$ on $D$;

end

return $f'$
We then extend query-synthesis-based augmentation method to multi-class classification, by the simple one-vs-one approach, which decomposes a multi-class task into multiple binary classification tasks. More specifically, if there are \( k_1 \) classes, then we can solve \( k_2 = k_1(k_1 - 1)/2 \) binary classifications instead, and use our active learning strategy to synthesize \( n_{\text{max}}/k_2 \) EEG epochs (where \( n_{\text{max}} \) is the maximum number of synthesized EEG epochs in one iteration) for each binary task.

### C. Adversarial Example Crafting for the Target Model

After training the substitute model, we can generate adversarial examples from it for the target model. Goodfellow et al. [15] proposed to construct adversarial perturbations in the following way:

\[
\delta = \varepsilon \cdot \text{sign}(\nabla_x J(\theta, x_i, y_i)),
\]

where \( \theta \) are the parameters of the target model \( f \), and \( J \) is the loss function. The main idea is to find an optimal max-norm perturbation \( \delta \) constrained by \( \varepsilon \) and \( l_\infty \)-norm is used as the distance metric.

Let \( \varepsilon = \varepsilon \) so that we can perturb \( x_i \) at the maximum extent. Then, the adversarial example \( x_i^\varepsilon \) can be re-expressed as:

\[
x_i^\varepsilon = x_i + \varepsilon \cdot \text{sign}(\nabla_x J(\theta, x_i, y_i)).
\]

UFGSM [22] is an unsupervised extension of FGSM, which replaces the label \( y_i \) in (4) by \( y_i' = f(x_i) \). Then, \( x_i^\varepsilon \) in UFGSM can be written as:

\[
x_i^\varepsilon = x_i + \varepsilon \cdot \text{sign}(\nabla_x J(\theta, x_i, y_i')).
\]

UFGSM was used in this paper to construct the adversarial examples.

### IV. EXPERIMENTS

This section presents the experimental results to demonstrate the effectiveness of the query-synthesis-based active learning strategy in transferability-based black-box attacks. We evaluate the vulnerability of three CNN classifiers in EEG-based BCIs.

#### A. Experimental Setup

The three BCI datasets, P300 evoked potentials (P300) [40], feedback error-related negativity (ERN)\(^1\) [41], and motor imagery (MI)\(^2\) [42], used in our recent study [22] were used again in this study. The data pre-processing steps were also identical. Three CNN classifiers, EEGNet [9], DeepCNN [10], and ShallowCNN [10] were used as the target models in our experiments, as in [22]. Adam optimizer [43], cross-entropy loss function, and early stopping were used in training. Moreover, we applied weights to different classes to address the class imbalance problem in P300 and ERN.

\(^1\)https://www.kaggle.com/c/inria-bci-challenge

\(^2\)http://www.bbci.de/competition/iv/

Raw classification accuracy (RCA) and balanced classification accuracy (BCA) were used to evaluate the attack performance, where RCA is the unweighted overall classification accuracy on the test set, and BCA is the average of the individual RCAs of different classes.

In order to simulate the black-box scenario, we partitioned the three datasets into two groups, as described in [22]: the larger group A of 7/14/7 subjects in P300/ERN/MI was used to simulate the unknown data from the target model, where 80% epochs were for training the target model and the remaining 20% for testing. The other smaller group B of 1/2/2 subjects was used to initialize set \( S_0 \) for training the substitute model. In this way, the 8/16/9 subjects in three datasets can be partitioned in 8/120/36 different ways. We performed both one-division and multi-division black-box attacks. In the one-division experiment, the attacks were repeated 5 times to reduce randomness. In the multi-division experiment, we repeated each division 5 times on P300 and only one time on ERN and MI; so, in total we had 40/120/36 evaluations on the three datasets, respectively.

#### B. Baseline

We first evaluated the baseline performance of the three CNN target models on the unperturbed EEG data, as shown in the first part of Table I. Because the MI dataset had 4 classes whereas P300 and ERN had only two, its RCAs and BCAs were much lower.

We then constructed a random perturbation \( \delta' \):

\[
\delta' = \varepsilon \cdot \text{sign}(N(0,1)),
\]

which has the same maximum amplitude \( \varepsilon \) as the adversarial perturbations, to verify the necessity of deliberately constructing the adversarial examples. The results are shown in the second part of Table I. It is obvious that the target models were robust to random noise, i.e., random noise cannot effectively perform adversarial attacks.

#### C. Attack Performance Comparison

We next compared our query-synthesis-based approach with the Jacobian-based method [22] in black-box attacks. \( \varepsilon = 0.1/0.1/0.05 \) on P300/ERN/MI were used to construct the adversarial examples after the substitute models were trained.

In one-division experiments, we randomly downsampled the data for each class according to the labels that the target model predicted at the first time to balance the classes in the initial dataset. We set the downsampled number in each class as 200 in P300 and ERN, and 100 in MI, to ensure that the size of the initial substitute model training set \( S_0 \) was 400. The initial substitute model was then trained on \( S_0 \) using both the Jacobian-based method and ours. \( \lambda = 0.5 \) and \( N = 1 \) were used in the Jacobian-based method, where \( N \) corresponds to \( N_{\text{max}} \) in our method. \( n_{\text{max}} = 200 \) and \( m = 10 \) were used in our method, so \( n_{\text{max}} = 2 \) should be used to get the same number of queries as in the Jacobian-based method. \( q = 1.0/0.8/0.8 \) on P300/ERN/MI were used in the experiments. They were determined such that the generated
EEG epochs had approximately the same magnitudes as those in the initial training set \( S_0 \).

In multi-division experiments, we did not limit the downsampled number to avoid wasting data, but always kept the same number of queries in the Jacobian-based method and ours, so \( n_{\text{max}} = \text{downsampling number} \) on P300 and ERN, \( n_{\text{max}} = 2 \times \text{downsampling number} \) on MI in our method, the other hyper-parameters were the same as those in one-division experiments. The attack results are shown in Table I. We can observe that:

1) Generally, the RCAs and BCAs of the target models after attacks were lower than the corresponding baselines, indicating the effectiveness of the transferability-based black-box attack methods. However, shallowCNN did not perform well in attacking EEGNet and deepCNN on P300 and ERN.

2) The RCAs and BCAs after query-synthesis-based black-box attacks were generally lower than the corresponding accuracy after Jacobian-based black-box attacks. For example, in one-division experiments in which only 400 queries were used for both methods, when the target model and the substitute model were EEGNet and DeepCNN respectively on P300, our method achieved 13% improvement over the Jacobian-based method.

3) Our proposed method was often much better than the Jacobian-based method when the substitute model and the target model had the same structure, which may due to the strong intra-transferability of the adversarial examples.

We compared the attack performance with different number of queries when the substitute model and the target model had the same architecture, i.e., both were EEGNet, DeepCNN or
ShallowCNN. 200 EEG epochs were used for initializing the substitute model, and the number of queries varied from 200 to 6,200. The corresponding RCAs/BCAs are shown in Fig. 1. Generally, both RCAs and BCAs decreased as the number of queries increased, which is intuitive. With the same number of queries, our method achieved lower RCAs and BCAs than the Jacobian-based method, i.e., it gave higher attack success rates. In other words, it can achieve a desired attack success rate with much fewer queries. For example, for EEGNet on P300, our method only needed 3,000 queries to achieve a comparable attack performance with the Jacobian-based method with 6,200 queries.

D. Characteristics of the Adversarial Examples

In this section, we analyzed the characteristics of the generated adversarial examples in both time domain and spectral domain. Consider one-division query-synthesis-based black-box attacks on ERN after 400 queries, when both the target model and the substitute model are DeepCNN. An example of the original EEG epoch (we only show the first 10 channels) and its corresponding adversarial epoch is shown in Fig. 2. They were almost completely overlapping in the time domain, which means the adversarial example is very difficult to be detected by human or a computer. We had similar observations on other datasets and from other classifiers.

Next, spectrogram analysis was used to further explore the characteristics of the adversarial examples. Fig. 3 shows the mean spectrogram of all original EEG epochs, of all adversarial examples whose predicted labels did not match the ground-truth labels of the original EEG epochs (successful attacks), and of the corresponding successful perturbations (x* – x), by using wavelet decomposition. The adversarial examples were crafted by query-synthesis-based black-box attacks on MI. Both the target and the substitute models
were DeepCNN. There was no significant difference in the
spectrograms of the original EEG epochs and the adversarial
eamples. The amplitudes of the mean spectrogram of the
adversarial perturbations were much smaller than those of
the original and adversarial examples, suggesting that our
generated adversarial examples are difficult to be detected by
spectrogram analysis. We had similar observations on other
datasets and from other classifiers.

V. CONCLUSION

In this paper, we have proposed a query-synthesis-based
active learning strategy for transferability-based black-box
attacks. It improves the query efficiency by actively synthe-
sizing EEG epochs scattering around the decision boundary
of the target model, and thus the trained substitute model can
better approximate the target model. We applied our method
to attacking three state-of-the-art deep learning classifiers in
EEG-based BCIs. Experiments demonstrated its effectiveness
and efficiency. With the same number of queries, it can achieve
better attack performance than the traditional Jacobian-based
method; or, in other words, our approaches needs a smaller
number of queries to achieve a desired attack success rate.

ACKNOWLEDGEMENT

This research was supported by the IEE Computational
Intelligence Society Graduate Student Research Grant, and the
National Natural Science Foundation of China (61873321).

REFERENCES

1–27.
a mental prosthesis utilizing event-related brain potentials,” Electroen-
ccephalography and Clinical Neurophysiology, vol. 70, no. 6, pp. 510–
boration effort,” IEEE Trans. on Human-Machine Systems, vol. 47, no. 4,
EEG headsets made easy: reducing offline calibration effort using active
weighted adaptation regularization,” IEEE Trans. on Neural Systems and
computer communication,” Proceedings of the IEEE, vol. 89, no. 7,
of stimulation methods used in SSVEP-based BCIs,” Computational
Intelligence and Neuroscience, p. 702357, 2010.
EEG-based brain-computer interfaces,” Journal of Neural Engineering,
K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball,
“Deep learning with convolutional neural networks for EEG decoding
and visualization,” Human Brain Mapping, vol. 38, no. 11, pp. 5391–
5420, 2017.
tions from EEG with deep recurrent-convolutional neural networks,” in
Proc. Int’l Conf. on Learning Representations, San Juan, Puerto Rico,
May 2016.
cation of EEG motor imagery signals,” Journal of Neural Engineering,
for online decoding of motor imagery movements from EEG signals,”
a crime: real and stealthy attacks on state-of-the-art face recognition,” in
adversarial examples,” in Proc. Int’l Conf. on Learning Representations,
Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” in
Proc. Int’l Conf. on Learning Representations, Banff, Canada, Apr.
2014.
the physical world,” in Proc. Int’l Conf. on Learning Representations,
adversarial examples,” in Proc. 35th Int’l Conf. on Machine Learning,
“Adversarial perturbations against deep neural networks for malware
https://arxiv.org/abs/1606.04435

Fig. 3. Mean spectrogram of the original EEG epochs, of all successful adversarial examples, and of the corresponding perturbations, in one-division
query-synthesis-based black-box attacks on the MI dataset. Channel $F_9$ was used.


